

NOVAS TECNOLOGIAS DE ANÁLISE DE DADOS PARA NOVAS AMEAÇAS



João Bastos Rocha (*)
Tenente-Coronel de Transmissões

RESUMO

Combater novas ameaças assimétricas como o terrorismo, o crime organizado e a lavagem de dinheiro, requer um sistema de Informações ágil que actue de forma activa e que faça uso das mais avançadas tecnologias de informação e comunicação.

As técnicas de *data mining* e análise automática de dados são ferramentas poderosas, que ajudam ao combate das novas ameaças, e que se encontram disponíveis para serem implementadas pelos Serviços de Informações. Todavia são ferramentas que também geram controvérsia e preocupações, pois apesar de fazerem uma análise de dados de forma fiável e eficaz, podem entrar em conflito com o direito à privacidade.

A adopção de sistemas de *data mining* pode ser a resposta a estes novos desafios. O presente estudo aborda as necessidades de implementação destes sistemas, as preocupações e receios que podem ser geradas e as soluções, ao nível da tecnologia, para os minimizar. Embora estas técnicas sejam muito poderosas, é um erro olhá-las como soluções completas para os problemas de Segurança, uma vez que é impossível eliminar completamente a subjectividade na análise da informação.

Palavras Chave: *Data Mining*, Ameaças assimétricas, bases de dados, privacidade.

(*) Professor das Unidades Curriculares de Programação de Sistemas e de Tática de Transmissões na Academia Militar

*Knock, knock.
"Who's there?"
"FBI. You're under arrest."
"But I Haven't done anything."
"You will if we don't arrest you," replied
Agent Smith of the Precrime Squad.¹*

INTRODUÇÃO

No período da guerra-fria era relativamente simples identificar as ameaças e as necessidades de Informações. Poderia ser difícil obter a informação necessária, mas o que se procurava era bem claro: conhecer as movimentações da outra superpotência, as suas possibilidades reais e as suas intenções. Actualmente as ameaças assimétricas são caracterizadas por colocarem graves escolhos aos Serviços de Informações. Ameaças como o terrorismo, o crime organizado altamente violento, o tráfico de armas de produtos radioactivos e de pessoas, são perpetradas por indivíduos sem escrúpulos, difíceis de identificar, organizadas em redes complexas e com o objectivo, que vai para além do meramente económico e político, e chega mesmo a ter em vista destruir o nosso modo de vida. O trabalho dos Serviços de Informações é colossal. Tem que encontrar estes criminosos num "oceano de ruído", compreender os seus padrões de actuação e desenvolver meios para prevenir as suas acções.

Creemos que a tecnologia pode dar um contributo decisivo neste árduo trabalho. É necessário encontrar novas formas mais inteligentes e eficientes de colecta e análise de dados, deve-se garimpar a informação para encontrar "pontos" que, depois de unidos, forneçam uma imagem fiável das actividades que procuramos. É imperativo transformar informação em conhecimento em tempo útil, criar normas legais de análise e disseminação da Informação para se obterem novas opções na prevenção das ameaças. As técnicas de *data mining* são uma ajuda fundamental que, garantindo a privacidade e a segurança, podem contribuir para uma melhor avaliação das ameaças.

O presente artigo pretende motivar e introduzir o uso dos sistemas de *data mining* como ferramenta fundamental para os serviços de Informações.

² Ver filme Relatório Minoritário, 20th Century Fox 2002.

1. NOVAS AMEAÇAS

Ameaças assimétricas apresentam-se hoje fruto da situação complexa em que o mundo vive. Após o desmoronamento da ordem imposta pela guerra fria assistiu-se à invasão do Kuwait, à desagregação da Jugoslávia, ao aparecimento dos "estados falhados", ao surgimento de ataques terroristas de dimensão trágica e sem racionalidade aparente. Os ataques de Nova York, Madrid, Bali ou Beslan mostram o horror de que os terroristas sem rosto são capazes por causas, muitas vezes, desconhecidas ou irracionais.

O ambiente estratégico que o mundo vive caracteriza-se pelo facto das *"ameaças, antes latentes, serem hoje bem reais, principalmente o terrorismo internacional e a proliferação de armas de destruição maciça. Elas são dirigidas contra os nossos concidadãos, contra os nossos valores, contra os nossos interesses, contra as nossas instituições democráticas."* (Raffarin, 2002, p. 5). As ameaças assimétricas usam métodos e meios não convencionais, que têm por objectivo circunscrever ou destruir as forças materiais e morais de um adversário, explorando todas as suas vulnerabilidades e fraquezas, incluindo as não militares, provocando efeitos potencialmente desproporcionais. Há uma grande dificuldade em conhecer o armamento de que as redes terroristas dispõem, pois a sua aquisição é efectuada ilegalmente. Por outro lado, estes grupos não se coíbem de usar meios não militares como arma e causar danos devastadores, ao nível moral e material.

O general Loureiro dos Santos considera que as novas ameaças são concretizadas por actores não clausewitzianos de natureza criminosa ou política. Os primeiros pretendem obter o máximo lucro sem quaisquer escrúpulos, usando o crime organizado, o tráfico de drogas e de pessoas e a corrupção estatal para atingir os seus fins. Por vezes, podem-se confundir com o Estado, mas não pretendem assumir o controlo e direcção política do Estado. Pretendem controlar o Estado na sombra, colocando políticos corruptos na sua chefia. *"Os segundos actores são clara e abertamente natureza política, embora se dissimulem sob outras capas, nomeadamente as étnicas e religiosas, e lançam mão, se o considerarem útil, de todas as actividades criminosas. Visam obter o poder político nacional ou global"*. (Santos, 2004, p. 201)

Notemos que estes actores trabalham muitas vezes em simbiose. Os que têm fins políticos usam os que têm fins económicos para obter fundos e estes usam técnicas de terror para multiplicar os seus fundos. São actores desterritorializados, organizados em redes, cada vez mais complexas e autónomas, difíceis de eliminar, pois sem grande relação hierárquica.

"Information analysis is the brain of homeland security. Used well, it can guide strategic, timely moves throughout our country and around the world. Done poorly, even armies of guards and analysts will be useless." (Baird, 2002, p. 6)

2. ANÁLISE DE AMEAÇAS

Desde sempre o homem quis prever o futuro, conhecer a acção do seu inimigo para lhe poder fazer face, seja negociando, seja preparando a sua defesa, ou atacando preventivamente. Os serviços de Informações têm acometida a tarefa de antever sobre quais as ameaças que se podem concretizar e qual o seu grau de perigosidade. É uma tarefa de previsão, nem sempre correcta porque além de se basear em factos conhecidos e na história, também se baseia na percepção do analista. É uma tarefa intelectual de inteligência e de estudo de comportamentos que se baseia em indicadores e avisos.

A forma de analisar e avaliar as ameaças tem evoluído ao longo dos tempos e depende fortemente do ambiente estratégico. Iremos tecer algumas considerações sobre indicadores e avisos, as formas tradicionais de avaliação de ameaças.

2.1 *Conceito de indicadores e avisos*

Os indicadores e avisos nos sistemas de Informações são usados para alertar os decisores políticos do início da crise de forma a poderem reagir em tempo útil. Um aviso prende-se com a tomada de decisões depois de se conhecer a informação disponível confirmando uma ameaça. Não é um acontecimento isolado no tempo, antes é um processo cíclico em que uma crise, um risco ou uma ameaça identificáveis são avaliados, dando origem à definição de um problema de alerta e ao estabelecimento de uma lista de indicadores decisivos. Em seguida, os indicadores decisivos são controlados em permanência e os sistemas de avaliação são actualizados.

Os indicadores são acções ou acontecimentos que ajudam a perceber se algum evento prejudicial pode surgir, são elementos de previsão. Consoante a situação vai evoluindo, podem-se criar indicadores decisivos ou cruciais, destinados a serem continuamente monitorizados, por representarem um indício significativo do que se está a passar. Estes devem ser definidos tão cedo quanto possível pois, se observados, transformam-se em indicações

claras do estado final de uma série de acontecimentos. Os indicadores cruciais são os acontecimentos que materializam a decisão do analista e podem, ou não, provocar a emissão de um aviso sobre a situação de crise. Os sistemas de indicadores e avisos pretendem detectar e relatar, em tempo oportuno, acontecimentos que ameacem uma entidade política de forma a permitir a tomada de decisão suportada em factos verificáveis. A implementação deste sistema é tanto mais difícil quanto mais complexo e volátil for o ambiente estratégico.

2.2 *Métodos de análise*

Durante os períodos das guerras clausewitzianas a avaliação da ameaça assentava na recolha de informação sobre as capacidades militares do inimigo e na sua determinação em empreender uma guerra. Os indicadores eram baseados na sua capacidade industrial, na indústria de guerra que implementava e na situação política que se vivia. A obtenção de Informações era efectuada por espões, meios diplomáticos, e, quando disponível, por meios tecnológicos avançados como satélites, fotografia aérea, entre outros. A recolha de informação sobre as capacidades militares do inimigo era crucial, era necessário conhecer o seu dispositivo, o número de homens, o seu armamento e a capacidade de produção caso um conflito deflagrasse. Por outro lado, era premente conhecer a situação política que se vivia, sendo esta, muitas vezes, a principal razão da concretização, ou não, da ameaça. (Cf. Herman, 2003, pp. 82 a 89). Por muito evoluídos que fossem os sistemas de obtenção de informação, aquela que sempre foi a preferida pelos serviços de Informações, foi a possibilidade de obter e decifrar mensagens entre as diversas organizações do adversário. Esta é a melhor forma de conhecer as intenções do inimigo, não sendo isenta de erros porque pode também ser uma forma de decepção.

Durante o período da guerra-fria, os indicadores eram baseados nos passos que o adversário deveria executar para preparar uma força militar. Eram indicadores mensuráveis, muito baseados em dados militares, embora não exclusivamente, que podiam antecipar qualquer agressão. O trabalho de análise era objectivo e visava recolher indicações suficientes de preparação de uma força militar que pudesse, com sucesso, invadir outro país. Os meios tecnológicos estavam vocacionados para a obtenção de informações sobre o desenvolvimento industrial de novos sistemas de armas. A HUMINT,

mais conhecida por espionagem ², visava recolher informações sobre os chefes militares e segredos de armamento, se possível roubar alguns sistemas de armas.

O método tradicional de avaliação das ameaças era baseado na sua probabilidade, perigosidade e possibilidade como factores de análise, e assentava em métodos indutivos e dedutivos, fortemente apoiado na experiência do analista, na história conhecida e na capacidade de recolha de informação crítica, efectuada, muitas vezes, de forma clandestina. (DeRosa, 2005, p. 5). A análise de informação consistia em efectuar inferências sobre os diversos indicadores observados e sobre as possibilidades do inimigo.

"For counterterrorism, we must be able to find a few small dots of data ³ in a sea of information and make a picture out of them"
(DeRosa, 2004, p. 5)

3. ANÁLISE DE DADOS COM DATA MINING

Após o que foi referido não só é claro que estamos perante um novo espectro de ameaças muito diferentes daquelas que se apresentaram aos Estados durante quase meio século de guerra fria, como também, na área da obtenção e análise de Informações, é necessário o emprego de novos meios e tecnologias de forma a perceber as novas ameaças e prever as acções de novos actores não estatais, que não se coíbem de utilizar meios não tradicionais.

A luta contra estas novas ameaças assenta em reduzidas peças de informação de diversas fontes que terão que se juntar para se fazer uma avaliação válida dessa informação. Os sistemas de *data mining* poderão ser uma ferramenta poderosa na descoberta de conhecimento num oceano de ruído.

² Para alguns autores os conceitos de HUMINT e de espionagem são diferentes. Aqui usamos a pirâmide de HUMINT de Michael Herman por a considerarmos como a mais relevante na sistematização dos conceitos. (2003, p. 61 a 66).

³ A terminologia inglesa adoptou a expressão "connecting the dots", unir os pontos, para classificar a obtenção de Informações referentes às novas ameaças. Esta expressão é uma analogia com os passatempos infantis em que unindo pontos numerados se obtém uma imagem com significado.

3.1 *Conceito de Data Mining*

Por *data mining* entende-se o processo de descobrir informações relevantes, como padrões, associações, mudanças, anomalias e estruturas, em grandes quantidades de informações armazenadas em bancos de dados, repositórios de dados ou outros mecanismos de guarda automática de informação. É um mecanismo de transformação de dados de baixo nível em informação de alto nível, ajudando no processo de tomada de decisão através do uso de algoritmos de exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações, ou ainda a comparar com padrões conhecidos.

Em suma, o *data mining* é um conjunto de procedimentos para extracção de padrões a partir de bancos de dados e insere-se no processo global de descobrimento de conhecimento útil em bases de dados, KDD (*Knowledge Discovery in Databases*). O KDD é a base de um sistema de informação inteligente. Compõe-se de um conjunto de passos a serem executados sobre uma, ou mais, base de dados inicial com o objectivo de extrair os conhecimentos desejados que possam estar contidos nessas bases de dados. O *data mining* é a parte principal do processo de KDD. (Cf. Gonçalves, 2003).

A análise de grandes volumes de informação não é nova. Tem sido usada pelo sector privado para a obtenção de informações e padrões de comportamento de clientes e pelo sector público na área da fraude e evasão fiscal. Em Espanha o projecto ZUJAR ⁴ aplica técnicas de *data mining* para detecção de evasão e fraude fiscal, a aplicação CORAL ⁵ usa estas técnicas para descobrir sistemas de lavagem de dinheiro, o Projecto ECHELON ⁶ usa-as para analisar comunicações ao nível mundial, o governo dos EUA utiliza *data mining* para identificar padrões de transferências de

⁴ ZUJAR; este projecto da administração tributária de Espanha está a ser implementado usando sistemas de data warehouse e data mining. Pretende-se encontrar todos os contribuintes que tenham padrões de comportamentos iguais aos que fogem ao pagamento de impostos. Informações mais detalhadas podem ser encontradas em: <http://www.dgci.min-financas.pt/ciat/DocsTecnicos/espanhol/1espana.doc>.

⁵ CORAL; Money Laundry Pattern Learning and Detection Using Data Mining Techniques, A aplicação tem uma versão de demonstração na Internet: <http://www.fortune.binghamton.edu/demo/CoralDemo.html>

⁶ ECHELON; o mais mediático e contestado programa dos EUA na área de escutas telefónicas, análise de correio electrónico e quaisquer telecomunicações. Estas escutas ocorrem em todo o mundo e usam técnicas de data mining para seleccionar as comunicações que devem ser alvo de escutas permanentes.

fundos internacionais que se assemelhem a lavagem de dinheiro do narcotráfico.

As técnicas de análise automática de dados ou de procura de padrões não substituem o analista de Informações, mas ajudam a libertá-lo de tarefas mecânicas de explorar grandes volumes de informação, para que ele se possa dedicar a assuntos que requeiram o seu julgamento. Estas técnicas também ajudam a priorizar a informação mais relevante e a eliminar aquela sem significado para o analista.

3.1.1 *Análise por assuntos ou ligações*

A análise de grandes volumes de informação é efectuada, tradicionalmente através de assuntos ou análise de ligações. Ou seja, quando um analista pretende conhecer as actividades de determinado indivíduo começa por procurar o seu cadastro, o seu registo criminal, as transacções bancárias que realiza, as deslocações e as pessoas com quem está ou esteve em contacto. Para cada informação que vá recolhendo pode, depois, ir acrescentando mais um elo numa cadeia de ligações que vai efectuando. Pode até ser ajudado por software poderoso, tipo "i2"⁷, que permite correlacionar factos e a descobrir ligações imperceptíveis à primeira vista. Esta é a actividade normal e mais usada pelos investigadores, quer no âmbito das polícias, quer no âmbito dos serviços de Informações. Ou seja, para se começar a obter e analisar informação sobre uma actividade ou uma pessoa começa-se por interrogar as bases de dados com o assunto ou nome da pessoa em questão. Por vezes, é possível também obter informação de outras actividades ilícitas praticadas por pessoas ou por organizações através do conjunto de ligações que vão sendo estabelecidas. Uma análise de ligações poderá ser efectuada com vários graus de separação e, assim, obter um conjunto de Informação que possa impedir que a ameaça que estamos a analisar se concretize.

⁷ i2 Software usado para encontrar relações entre pessoas, organizações ou factos com apresentação de forma tridimensional, que seriam muito difíceis de encontrar com os métodos de análise tradicionais. É usado pelos serviços de Informações e pela investigação criminal. Em Portugal o BISM tem usado, experimentalmente, este software, com resultados excelentes. Uma demonstração pode ser vista em www.i2.co.uk.

Uma análise dos ataques de 11 de Setembro, efectuada pela *Markle Foundation Task Force*, mostra como é possível extrair ligações complexas e obter informação relevante sobre os planos dos terroristas. Se tivesse sido possível integrar, ao mais alto nível, toda a informação disponível, tais como listas de pessoas sob suspeita, registos de reservas de voos e dados de residências, entre outros, os terroristas poderiam ter sido identificados a tempo de se poder iniciar uma investigação. (Baird, 2002, p. 28)

3.1.2 *Análise de padrões*

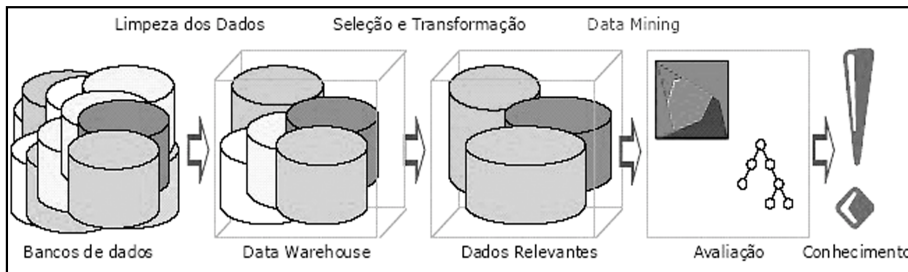
A análise de padrões é menos objectiva e de efectuação mais complexa. Não se procura um suspeito, mas procurar-se encontrar padrões de comportamento, que se conheçam como ilícitos, e depois encontrar quem actua segundo esses padrões. Ao contrário da pesquisa por assuntos, a pesquisa por padrões não necessita de uma pista inicial. Apenas se necessita de conhecer um padrão de actividade ilegal. Pesquisas baseadas em padrões podem servir para investigar, sistemas conhecidos de lavagem de dinheiro, actividade terrorista adormecida, crime organizado, tráfico de seres humanos e narcotráfico. Por exemplo, se uma rede de lavagem de dinheiro com ligações a uma rede terrorista for identificada, descobrindo o padrão de actuação, pode-se tentar encontrar o mesmo padrão numa base de dados. Se o padrão for encontrado teremos que prosseguir as investigações reduzindo o universo de possíveis suspeitos. À medida que o processo for evoluindo podemos chegar a um conjunto de suspeitos que poderão ficar sob vigilância para se aferir da actividade criminosa ou não. O mesmo se passa com a análise de tráfego na Internet ou de actividades terroristas. Se procuramos terroristas que possam efectuar um ataque com camiões carregados de explosivos, podemos analisar o aluguer de camiões, em conjunto com a aquisição de explosivos para actividades legais e o alojamento em hotéis e a recepção de mensagens de correio electrónico ou correio suspeito. Padrões de envio de correio electrónico, de conversas telefónicas, números de contas, hotéis registados e pagos em dinheiro, podem indicar onde se encontram os líderes e como enviam mensagens para os seus operacionais.

Estas ferramentas de análise são largamente usadas no meio comercial para obtenção de padrões de consumo, promovendo produtos que se prevê ter interesse para o cliente. Em Portugal, as cadeias de supermercados com cartão de cliente são as que mais usam estas técnicas, as instituições bancárias empregam-nas para a concessão de crédito, e as redes internacionais de vendas pela Internet usam-nas para aumentar os lucros e fidelizar clientes ⁸.

É necessário termos um padrão de comportamento para podermos detectar outros. Este padrão ou é conhecido de grupos que já o usaram ou pode ser inferido, sendo, neste caso, a possibilidade de erro muito maior.

3.2 O Processo

Na figura seguinte apresenta-se o processo de extracção de conhecimento a partir de bases de dados. Vamos apresentar o que consideramos como importante conhecer para permitir a tomada de decisão de emprego deste sistema.



Fonte: www.lac.inpe.br/eventos/downloads/data.pdf

Figura 1 - Processo de extracção de conhecimento

⁸ Isto pode ser verificado na aquisição de livros na Amazon.com. Sempre que o cliente adquire um livro é-lhe proposto a aquisição de outros títulos, informando, claramente que quem compra determinado livro "costuma" comprar os títulos propostos. As empresas que usam a Internet para vender os seus produtos e a própria disposição dos produtos nos expositores obedecem a estudos de padrões de consumo de forma induzir o cliente a adquirir o máximo. Estes padrões de consumo são analisados usando técnicas de *data mining* e de comportamento social.

3.2.1 *Recolha e processamento de dados*

O primeiro passo consiste em identificar, recolher e processar os dados que, posteriormente, serão analisados. Este é uma etapa fulcral do sistema, pois pode tornar o sistema insignificante se não tiver os dados correctos ou se o "lixo" for tanto que torna inoperacional o funcionamento da pesquisa. A eliminação de registos errados de uma base de dados é uma tarefa muito complexa.

Os dados seleccionados na etapa anterior são então combinados numa base de dados chamada de *data warehouse*, para depois ser pesquisada. Esta *data warehouse* pode ser física ou virtual. No primeiro caso a informação é recolhida a partir dos vários organismos que se encontram legalmente possibilitados para recolher a informação criando-se uma base de dados centralizada. No segundo caso apenas é estabelecido um conjunto de ligações que permitam a criação de bases de dados distribuídas. Nas bases de dados centralizadas temos a vantagem da facilidade de análise da informação e como desvantagem a manutenção e actualização dos dados. Nas bases de dados distribuídas acontece o contrário.

Apesar de esta ser a forma mais eficaz de trabalhar, se existirem constrangimentos legais ou de acesso podemos aplicar uma análise de padrões sobre uma base de dados tal como se encontra. Se for necessário garantir a privacidade no uso da informação poder-se-á criar uma arquitectura distribuída que permita diferentes acessos e normas de aplicação diferentes consoante se for encontrando transacções suspeitas. Este processo dificulta o uso da base de dados para outros fins que não os previstos inicialmente.

O passo seguinte consiste na transformação e normalização de dados. Este tratamento de informação é fundamental para se poder obter um conjunto de dados relevantes. Por exemplo, numa base de dados os nomes podem estar completos ou abreviados. Tó Zé pode significar António José ou Luís pode ser nome próprio ou apelido. Também as moradas são escritas de diversas formas. Há a Rua da Liberdade e a Avenida da Liberdade, R/C D pode significar Rés-do-chão direito ou fracção D. Em todas as bases de dados, e porque não há normalização entre entidades,

eventualmente poderá existir dentro de uma entidade ⁹, há diferentes formas de representar a informação. Ou seja, a informação deverá ser tratada para se poder ter dados relevantes e se aplicar o processo de *data mining*.

3.2.2 Modelos de pesquisa

Como foi referido, na análise de padrões é necessário a criação de um modelo que se possa aplicar aos dados. De uma forma genérica há dois processos ou métodos de *data mining*: de cima para baixo (*top-down*) ou de baixo para cima (*bottom-up*) ¹⁰.

A abordagem de cima para baixo começa com uma hipótese e procura validá-la. A hipótese pode ser elaborada tendo, inicialmente, pesquisado os dados usando a técnica de baixo para cima ou desenvolvida com conhecimento de um padrão real. A experiência ou dados dos serviços de Informações podem, e devem, ser a fonte destas hipóteses que serão aplicados aos dados. A determinação desta hipótese ou padrão é um trabalho que exige pessoal muito especializado e treinado.

O método de baixo para cima analisa os dados e extrai padrões ou anomalias que indiquem certo comportamento. Este método pode ser directo ou supervisionado (quando se tem alguma ideia do que se pretende encontrar) ou indirecto ou não supervisionado (quando não se tem ideia do que se procura). (Taipale, 2003, p. 30). Este método cria, a partir de um grupo, a identificação de um subgrupo de pessoas que usam um determinado padrão que depois será validado por fontes de Informação tradicionais ¹¹.

⁹ Mesmo em entidades onde há normas difundidas para a recolha de informação em bases de dados os erros são comuns. Os sistemas mais fiáveis são aqueles que obrigam o utilizador a escolher uma opção e que implementam códigos de segurança nos números identificadores. Um exemplo acontece com o número de contribuinte em que o último algarismo é um código de segurança que impossibilita o operador de registo de dados de digitar mal o número.

¹⁰ Apresentamos a terminologia inglesa porque é recorrente o seu emprego, nas obras escritas em português.

¹¹ A esta técnica chama-se de "*clustering*", agrupamento. São criados grupos de padrão comportamental perfeitamente definidos e conhecidos. Os novos dados são colocados na matriz de avaliação e é calculada a menor distância a cada grupo. O novo registo é então associado a esse grupo e determinada a margem de erro. Esta técnica é usada pelas instituições bancárias para a concessão de crédito e negociação da taxa de juro. (Taipale, 2003, p. 29 e seguintes.)

Os sistemas de *data mining* só poderão gerar conhecimento válido se o modelo for válido. Como os padrões que se procuram dizem, normalmente, respeito a pessoas que usam métodos evasivos, que tentam esconder e apagar todo o rasto das suas actividades quotidianas, a possibilidade de existência de falhas na pesquisa de informação é muito grande. Daí a necessidade de avaliar toda a informação disponível ao mais alto nível. Dada a exiguidade de informação e o "ruído" que esta possa ter, todos os aspectos são relevantes. Os EUA usam este sistema no programa, *Terrorism Information Awareness* (TIA) da *Defense Advanced Research Projects Agency* (DARPA) que engloba todas as fontes de dados do governo dos EUA. Este projecto baseia-se na pesquisa de informação através de *data mining*. Na ideia de John Poindexter, director do projecto TIA, os "EUA tornam-se assim, muito mais eficientes na pesquisa e geração de informação tornando-a disponível para análise, convertendo-a em conhecimento para fazer face a novas ameaças. Os EUA devem partilhar informação entre todas as agências e criar equipas de apoio de elevado desempenho operando nas margens das organizações terroristas". (Poindexter, 2002, p. 1) ¹². Outro projecto adoptado pelos EUA, usando sistemas de *data mining*, é o *Computer Assisted Passenger Pre-Screening* (CAPPS II), desenvolvido pela Agência de Segurança de Transportes após o 11 de Setembro. Na efectivação de uma reserva de um voo, o CAPPS II tem como objectivo fornecer à Agência de Segurança de Transportes, o nome do passageiro a sua morada, data de nascimento e o número de telefone de contacto o que permite efectuar uma verificação da identidade do passageiro e uma avaliação de risco de terrorismo ¹³.

¹² Este projecto começou por ser conhecido como Total Information Awareness. Este programa foi cancelado pelo Congresso dos EUA, em finais de 2002, e obrigado a ser integrado nos projectos da DARPA de forma a evitar a possibilidade de abusos sobre as liberdades dos cidadãos. Este programa continua a ser desenvolvido, em moldes relativamente diferentes, nos programas do Information Awareness Office (IAO) da DARPA. (Taipale, 2003, p. 19).

¹³ A consulta da Agência de Segurança de Transportes dos EUA permite que se confira as diferentes fases de registo de viajantes que pretendem entrar no país. Mais recentemente os EUA desenvolveram aplicações de registo, via Internet, obrigatório para todos os passageiros que pretendam entrar nos EUA ao abrigo do programa de isenção de vistos.

Estes sistemas estão implementados e a funcionar. São uma forma poderosa de avaliação da ameaça com tempo muito reduzido. Porém, como veremos na secção 3 deste capítulo, a utilização destes sistemas nem sempre é consensual, podem ser utilizados como uma forte ameaça às liberdades e privacidade dos cidadãos. Nos EUA há uma grande controvérsia com estes projectos tendo sido, por várias vezes, postos em causa pelo Congresso ¹⁴.

3.2.3 *Avaliação e tomada de decisão*

O processo de aplicação de técnicas de KDD a um conjunto de bases de dados termina com a avaliação dos resultados obtidos e com as decisões que poderão ter que ser tomadas sobre o emprego dos resultados. No contexto das políticas dos Serviços de Informações existem um conjunto de restrições de ordem legal que não podem ser omitidos. Apesar de nas organizações comerciais se usarem técnicas de análise de padrões de forma automática, nos serviços governamentais o seu emprego deve ser parcimonioso. Estas ferramentas deverão destinar-se à obtenção de informação inicial que será depois confirmada, ou não, pelos analistas de Informações. A avaliação dos resultados finais deverá ser sempre da responsabilidade de uma pessoa certificada para aceder a informação delicada.

No patamar de tomada de decisão dever-se-á ter em consideração os aspectos legais de uso da Informação, possibilidade de comunicação a outras forças policiais ou judiciais e ainda a possibilidade de extravio dessa Informação. Isto é muito relevante quando se procuram padrões muito difíceis de provar junto do poder judicial, antes de a acção acontecer. Por exemplo, se procurarmos padrões de lavagem de dinheiro e obtivermos sucesso, em teoria será possível descobrir a pessoa que efectuou e accionar os mecanismos legais necessários. No caso de evitar um atentado terrorista o problema é muito maior. De facto, antes de o atentado se perpetrar estamos no campo das intenções. Estas ferramentas podem detectar um padrão de

¹⁴ Na Internet estão disponíveis vários sítios que levantam objecção aos programas de obtenção de Informações acerca dos cidadãos. Alguns autores receiam mesmo que uma centralização muito grande da informação possa ter efeitos contrários aos que se anunciam.

comportamento que se conhece como o tido pelos terroristas e não ser terrorista. É difícil provar que uma pessoa tem em mente a execução de atentado terrorista se não existirem provas materiais, com planos, explosivos, mensagens, que o suportem.

A avaliação e uso de Informação obtida por meios de descoberta de conhecimento deve ser efectuada por pessoa habilitada e de confiança, sujeita a controlos rigorosos, e sempre que possível, usando algoritmos de ocultação de identidade das pessoas envolvidas.

3.3 *Riscos*

As ferramentas apresentadas são formas poderosas de descoberta de conhecimento que tenta ser escondido pelos seus autores. Como ferramenta poderosa na área dos serviços de Informações, acarreta sempre uma suspeita de uso abusivo, tema este sempre presente no debate sobre o seu emprego ¹⁵. De facto apresentamos, de seguida, os riscos que se correm com a utilização destas técnicas. Na secção seguinte apresentaremos também o que pode ser desenvolvido, ao nível da tecnologia, para obviar estes riscos.

3.3.1 *O fim da "privacidade"?*

A privacidade das pessoas é um bem que, nas sociedades ocidentais, é muito valorizado, sendo considerado um dos valores que define a liberdade da pessoa e das sociedades ao ponto de consideramos como evoluídas as sociedades que respeitam a privacidade dos seus

¹⁵ A UE mandou uma comissão do Parlamento europeu para "confirmar a existência do sistema de interceptação de comunicações conhecido por ECHELON, cujo funcionamento é descrito no relatório STOA sobre o desenvolvimento da tecnologia de vigilância e riscos de abuso de informações económicas; verificar a compatibilidade de tal sistema com o direito comunitário, designadamente com o artigo 286.º do Tratado CE, com as Directivas 95/46/CE e 97/66/CE, e ainda com o n.º 2 do artigo 6.º do Tratado UE à luz das seguintes questões:

- Os direitos dos cidadãos europeus encontram-se protegidos das actividades dos serviços secretos?
 - A criptagem constitui uma protecção adequada e suficiente para garantir a defesa da vida privada dos cidadãos, ou deverão ser adoptadas medidas complementares e, em caso afirmativo, que tipo de medidas? - De que modo poderão as Instituições da UE ser alertadas para os riscos decorrentes de tais actividades, e que medidas poderão ser adoptadas?
 - Verificar se a interceptação de informações a nível mundial constitui um risco para a indústria europeia;
 - Formular, eventualmente, propostas de iniciativas políticas e legislativas."
- (Schmid, 2001, p. 23)

cidadãos. Porém, a privacidade é um conceito subjectivo, depende da percepção de cada cidadão, da sua cultura, das convicções religiosas, entre muitos outros factores.

A privacidade e a segurança são conceitos que, muitas vezes, estão em contradição e requerem um tratamento equilibrado, devendo ser consideradas duas obrigações complementares do Estado junto dos seus cidadãos.

O uso de novas tecnologias poderá registar dados de diversa ordem da pessoa. O cartão do cidadão pode conter dados que, para alguns fundamentalistas da privacidade, podem ser considerados uma invasão à sua privacidade. Os sistemas de registo em hotéis, passagens de avião, o registo de posições de telemóvel, o registo das compras através de cartão de crédito, podem ser consideradas uma invasão na vida privada, mas nem por isso, podem deixar de ser implementadas. Todas estas formas de registo existem, são necessárias ao bom funcionamento destes sistemas e continuarão a existir, no actual estado de desenvolvimento tecnológico, tendendo mesmo a generalizar-se a outros aspectos da vida.

Os serviços de Informações não necessitam de recolher mais informação sobre as pessoas para, "ligando os pontos", a poder analisar e transformar em conhecimento. (Taipale, 2003, pp. 50 e ss.). Necessitam sim de autorização legal e de implementar mecanismos de controlo para poder relacionar um enorme volume de informação, que se encontra disperso e sujeito a legislação específica, de forma a obter Informação relevante em áreas muito sensíveis como o terrorismo, crime organizado, lavagem de dinheiro, tráfico de seres humanos, crimes ambientais.

Reconhecemos que o acesso a um volume de informação tão diverso, não está isento de riscos de uso abusivo e de pesquisas da vida privada de "pessoas públicas", podendo, no limite, ser usado com fins ilegais de manutenção ou derrube de governos.

3.3.2 *Falsos positivos e falsos negativos*

Um falso positivo acontece quando o processo reporta, de forma incorrecta, que encontrou o padrão que se procurava. Nos falsos negativos o processo indica o contrário, que não encontrou qualquer padrão, também de forma incorrecta.

Como o *data mining* é um processo de procura de padrões, não existem mecanismos que impeçam o sistema de reportar um padrão de comportamento típico de terroristas, num cidadão sem qualquer ligação a grupos criminosos. Outro problema pode acontecer com todos aqueles que têm ligações com os grupos criminosos fruto da sua actividade quotidiana. Se alguém arrendar casas ou alugar carros a terroristas, pode ser objecto de pesquisa sem qualquer fundamento. Aqui coloca-se um dos maiores desafios ao sistema: reduzir ao máximo a ocorrência de falsos positivos.

Os falsos negativos, por outro lado, não causam tantas preocupações de legalidade. De facto se o sistema indica que não encontrou nenhum padrão, os serviços não vão colocar ninguém sob vigilância. Não deixa de ser um problema e um desafio ao sistema evitar que existam falsos negativos, pois pode comprometer o seu funcionamento e a segurança do país.

Como já foi referido, estas ferramentas devem ser consideradas como poderosos meios de ajuda aos analistas e investigadores dos serviços encarregados da Segurança e da Defesa Nacional, produzindo pistas que permitam uma investigação posterior e uma confirmação, ou não, por meios tradicionais. (DeRosa, 2004, p. 15). Neste caso os falsos positivos podem ser descobertos antes de terem um impacto significativo na vida das pessoas. Todavia, se forem disseminados por outros serviços ou forças policiais, sem serem devidamente confirmados, pode ser muito difícil evitar causar danos às pessoas, contribuindo para o emprego de recursos de pesquisa de Informações em situações erróneas, causando mal estar nos agentes do sistema.

3.3.3 *Uso inadequado da Informação*

A concentração da informação, seja em *data warehouses* físicas ou virtuais, é sempre susceptível de uso abusivo por parte de pessoas ou entidades que lhe tenham acesso. Pode existir a tentação de se criar um "*big brother*", na expressão Orwelliana, que pesquise informações ou padrões de comportamento que não põem em causa segurança nacional, antes favorecem a acção dos governos, grupos de pressão ou, ainda, a chantagem sobre certas pessoas.

Os serviços de Informações são uma das formas de poder do Estado e, como tal, devem ser usados com resguardo da salvaguarda da legalidade, necessidade e pelo respeito das liberdades e garantias dos cidadãos.

Como a história mostra, num extremo, os Serviços de Informação podem constituir-se como centros de poder dentro do Estado, independentes do sistema político, enquanto, noutro extremo, podem servir o aparelho governativo para se perpetuar no poder ¹⁶.

Devemos realçar que os sistemas de descoberta de conhecimento não pretendem recolher mais informação, antes dispor da informação existente actualizada e em tempo, tendo objectivos determinados à priori e serem supervisionados pelo poder judicial. Assim, o uso indevido de Informação pode acontecer, não porque se recolhe mais informação sobre factos que afectem a vida privada das pessoas, mas porque os agentes intervenientes no processo usam a informação cruzada possibilitando o conhecimento de informações não relevantes para os objectivos iniciais, mas tentadora nos planos políticos, financeiros ou até criminais. Há sempre a tentação de expansão do sistema a áreas diferentes das inicialmente previstas.

O controlo e supervisão do uso destas ferramentas deve ser efectuado a vários níveis, interno e externo. Como veremos há algumas respostas, baseadas nas tecnologias mais recentes que podem ser implementadas, para atenuar os riscos apresentados.

3.4 *Atenuar os riscos*

As formas de minimizar os riscos enumerados na secção anterior devem ser baseadas na tecnologia, na definição de um quadro legal coerente, na selecção correcta dos intervenientes no processo e na protecção física das instalações e dos dados.

Neste trabalho limitar-nos-emos a tecer considerações baseadas na tecnologia abordando as áreas em desenvolvimento que permitem o tratamento de

¹⁶ Mesmo nos EUA, onde há uma grande tradição democrática, houve tentativas de uso dos Serviços de Informação para apoio das actividades de obtenção de Informações de "adversários" políticos sendo o caso Watergate, que levou à resignação do presidente Nixon, o mais emblemático.

informação de forma anónima, a redução de falsos positivos e tecnologias que permitam auditar e criar regras de uso e processamento de bases de dados. Todos estes subsistemas encontram-se já implementados, com diferentes graus de desenvolvimento, nas modernas bases de dados comerciais.

A primeira área de desenvolvimento tecnológico é a do tratamento anónimo de dados, para mascarar a informação de identificação do registo, de forma aos analistas poderem efectuar as suas pesquisas sem acederem à identidade real das pessoas. Um exemplo é o que ocorre com um aluno que executa um exame e a sua correcção é efectuada apenas conhecendo um número de teste. Claro que quando falamos em informações provenientes de múltiplas bases de dados o problema torna-se mais difícil de resolver. O primeiro passo a implementar consiste no mascaramento de elementos da identidade - tais como, nomes, endereços, números de bilhete de identidade, carta de condução, entre outros - através de códigos. Isto não resolve o problema por completo. Pode-se inferir a identidade de uma pessoa através do sexo, data de nascimento, local de trabalho, função que desempenha, etc. Para obviar este problema tem-se desenvolvido algoritmos de protecção de privacidade dos quais se destaca o "K-anonymity", desenvolvido por Latanya Sweeney na Carnegie Mellon University, EUA (Sweeney, 2002, pp. 557 a 570)¹⁷. Este algoritmo permite evitar que se tenha acesso a dados que identifiquem uma pessoa dentro de um grupo de K-1 outros indivíduos. Por exemplo se K for definido como 10000, só podemos inferir identidades num grupo de 10000 pessoas. (DeRosa, 2004, p. 18). O K é assim definido de acordo com os decisores políticos e com o decorrer da investigação. Quando se começa a procurar um padrão, K deve ter um valor muito elevado, à medida que se vão eliminando registos o K deve ir diminuindo de forma a chegar a um ponto em que existe um número de suspeitos passível de poder ser investigado por métodos tradicionais. Este método é conhecido como revelação selectiva, (Taipale, 2003, pp. 74 e 79) em que os investigadores vão tendo conhecimento das identidades à medida que a investigação vai progredindo. A autorização

¹⁷ O trabalho de Latanya Sweeney tem sido empregue pelo DARPA no seu programa de TIA. Há outras formas de protecção de dados desenvolvidos para a manipulação de dados usando o KDD. A discussão destes métodos pode ser acompanhada no sítio internet da fundação Markle.

concedida aos investigadores deve partir das entidades competentes, e as identidades só devem ser reveladas quando existirem certezas razoáveis sobre as actividades ilegais que se procuram.

A resolução de falsos positivos está directamente relacionada com o primeiro passo do KDD, em que a qualidade dos dados e a quantidade de "ruído" existente pode provocar distorções na pesquisa de padrões. Um dos programas mais complexos que está a ser desenvolvido em diferentes universidades e no DARPA é a criação de sistemas inteligentes que possam correlacionar dados de diferentes bases de dados sem duplicação de registos fruto da falta de normalização existente. Uma solução para o problema da normalização seria a criação de uma norma que implementasse um conjunto de procedimentos de recolha de informação. Este processo é de implementação relativamente demorada, tem problemas de natureza legal e os resultados seriam sempre medíocres pois muitos dos erros de colecta de informação provêm dos operadores de registo de dados.

A terceira área de desenvolvimento prende-se com os sistemas de auditoria. O controlo de acessos aos dados, em grandes bases de dados, é um processo comum na maioria das empresas comerciais, e é obrigatória a sua implementação nas bases de dados dos mercados financeiros. O sistema de auditoria permite conhecer quem acede aos dados, quais as pesquisas efectuadas, as horas a que o sistema recebeu ou forneceu informação e de onde foi acedido. Por esta razão, auditar dados pode ser também uma fonte de informação, pois pode-se conhecer as investigações em curso, quem as está executar, requerendo que os auditores sejam pessoas de idoneidade a toda a prova. É um mecanismo poderoso e eficaz de controlo da informação, requerendo o consumo de grandes recursos computacionais e humanos.

O último desenvolvimento tecnológico que vamos abordar prende-se com a implementação de regras de acesso e pesquisa, que permitam a protecção da privacidade. Este sistema apoia-se em dois elementos. Primeiro, as pesquisas são efectuadas de acordo com as permissões do utilizador ou da entidade que efectua a pesquisa, através de agentes inteligentes que são agregados à pergunta. Ou seja, uma pesquisa efectuada de acordo com uma autorização atribuída a um utilizador pode ter permissões diferentes da mesma pergunta que tenha autorização atribuída por uma entidade jurídica. Em segundo lugar, os dados podem ser etiquetados com metadados (dados sobre os dados), descrevendo como podem ser processados.

Assim, mesmo que os dados sejam copiados para uma base de dados diferente, mantêm as regras pelas quais podem ser processados garantindo a sua protecção inicial. O processamento de dados baseados em regras assenta em ferramentas de etiquetagem e filtragem de dados ¹⁸.

3.5 *O estado da arte*

Analisámos um sistema de pesquisa de conhecimento baseado na mais moderna tecnologia de bases de dados e nos avanços matemáticos na área da teoria dos conjuntos. Estes sistemas têm sido desenvolvidos comercialmente pela indústria da publicidade e nas pesquisas de mercado. O grande objectivo é conhecer padrões de consumo e padrões de comportamento face às promoções de vendas.

O desenvolvimento das ferramentas de KDD permitiu o seu emprego em áreas tão diversas como a medicina, na banca, nas telecomunicações, nas bibliotecas, etc. Actualmente estes processos estão a ser usados para as empresas obterem conhecimento do mercado e dos seus concorrentes directos e é também empregue nas áreas de Segurança e Defesa, como temos visto. Os EUA lideram as pesquisas efectuadas neste domínio, mas o Reino Unido, a Austrália, o Canadá, a França, a Espanha entre outros países desenvolvidos, têm programas com grande apoio das TIC, nomeadamente sistemas de descoberta de conhecimentos em bases de dados.

Ao nível tecnológico os principais fabricantes de bases de dados, Oracle, IBM, Sybase, implementaram mecanismos de *data mining* embebidos nos seus motores de pesquisa comerciais e têm sistemas que permitem auditar o seu acesso. O desenvolvimento matemático na área da teoria de conjuntos, da estatística e da algoritmia, permite-nos crer que é possível efectuar buscas de padrões complexos sem invadir a privacidade das pessoas e sem a criação de falsos positivos que tornem impossível a implementação do sistema.

¹⁸ O sistema de filtragem de dados DCS-100 ("Carnivore") é um sistema de filtragem analítica de dados desenhado para examinar tráfego de correio electrónico e só recolher o material que está autorizado. "O Carnívoro fornece ao FBI a possibilidade de interceptar e recolher comunicações que se encontrem sob a alçada da lei, enquanto ignora as comunicações que não são autorizadas interceptar. Funciona como os "sniffers" comerciais e outra ferramentas de diagnóstico de rede usadas pelos fornecedores de serviço de Internet". (Taipale, 2003, p. 78). Tradução do autor.

A formação dos utilizadores de um sistema de KDD é complexa. Deve promover um conhecimento profundo de pesquisas em bases de dados relacionais e orientadas a objectos, o conhecimento da legislação que lhe permita aceder a dados e ter formação na área de análise de Informações. Cremos que o ideal será formar equipas pluridisciplinares com uma forte componente tecnológica.

CONCLUSÕES

Em todos os tempos se procurou conhecer as ameaças à segurança dos povos e os modos de evitar que se concretizassem. Até há bem pouco tempo as ameaças eram identificadas como entidades precisas – Estados, organizações mais ou menos institucionalizadas e reconhecidas – e por isso era para elas que se dirigiam as investigações e desenvolvimentos técnicos na recolha de informação.

Com o fim da guerra fria e o advento do século XXI as ameaças mudaram, são agora mais difusas na sua identidade e, como tal, mais difíceis de identificar e, por isso, de estudar.

Apesar de continuar a ser um meio eficaz e necessário, o emprego de agentes infiltrados nas organizações terroristas e de crime organizado transnacional é muito difícil de concretizar. É um processo muito demorado, com resultados a longo prazo pelo que, os Estados têm de lançar mão de outros meios de obter Informações credíveis, relevantes e oportunas que lhe permitam garantir a segurança dos seus cidadãos face ao terrorismo, ao narcotráfico, à proliferação de ADM, à fraude económica e fiscal, aos ataques aos sistemas informáticos e outras ameaças que, mercê da globalização, chegam hoje a todos os recantos do planeta.

Um dos meios mais poderoso de armazenagem de informação está hoje disponível nas imensas bases de dados e a sua análise, pelo processo de *data mining*, poderá tornar essa informação altamente eficaz na identificação de ameaças latentes ou declaradas. Através do estudo de padrões e assuntos, por pessoal especializado, podem inferir-se situações de ameaça à Segurança e apresentar às instâncias de decisão dados precisos sob quem deve ser considerado suspeito e ser colocado sob vigilância, ou que actividades devem ser colocadas sob investigação por parte dos meios convencionais de vigilância.

É claro que estamos perante processos complexos e que não estão isentos de erro, mas que, face à escassa e difícil recolha de informação sobre as ameaças actuais, é um bom instrumento de trabalho que está ser posto ao serviço por inúmeros Estados.

BIBLIOGRAFIA

BAIRD, Zoë, et al., (2002). *Protecting America's Freedom in the Information Age*, in vários, Markle Foundation Task Force, Nova York, Internet: http://www.markletaskforce.org/report1_overview.html, consultado em 28 de Janeiro 2009.

DeROSA, Mary, (2004). *Data Mining and Data Analysis for Counterterrorism*, in Vários, Ed. Center for strategic and International Studies, Washington, D. C. Internet: www.csis.org, consultado em 28 de Janeiro 2009.

FAYYAD, Usama M. et al., (1996). *Advances in Knowledge Discovery and Data Mining*, ed. MIT, Massachusetts, USA.

GONÇALVES, Rodrigo e Hillesheim, (2003). *Sistemas de Informação inteligentes*, UFSC, Santa Catarina, 2003, Internet: www.inf.ufsc.br/~rodrigog/free/TranspIA.pdf, consultado em 6 de Setembro 2004.

GROTH, Robert, (1997). *Data Mining: A Hands-On Approach for Business professionals*, Ed. Prentice-Hall, New Jersey, USA

HERMAN, Michael, (2003). *Intelligence Power in Peace and War*, 1.^a Edição, Ed. Cambridge University Press, Cambridge.

POINDEXTER, John, (2002). *Overview of the Information awareness Office*, Conferência DARPAtech. Internet: www.darpa.mil, consultado em 28 de Janeiro 2009.

RAFFARIN, Jean-Pierre, (2002). *La politique de défense de la France*, in Défense Nationale, Paris.

SANTOS, José Alberto Loureiro dos, (2004). *Convulsões: Ano III da "Guerra" ao terrorismo - Reflexões sobre Estratégia IV*, 5.^a ed., Ed. Publicações Europa-América, Mem Martins.

SCHMID, Gerhard, (2001). Comissão Temporária sobre o Sistema de Intercepção ECHELON, *Relatório sobre a existência de um sistema global de intercepção de comunicações privadas e económicas (sistema de intercepção .ECHELON.)*, in Vários, Parlamento europeu, N.º PE 305.391. Internet: <http://www.europarl.europa.eu>, consultado em 28 de Janeiro de 2009.

SWEENEY, Latanya, (2002). "k-Anonymity: A Model for Protecting Privacy", in Vários, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems*, Internet: <http://privacy.cs.cmu.edu/people/sweeney/kanonymity.pdf>, consultado em 28 de Janeiro de 2009.

TAIPAILE, K. A. (2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data", in Vários, *Science and Technology Law Review*, Columbia, Vol. V. Internet: www.stlr.org, consultado em 28 de Janeiro 2009

